education sciences

MDPI

Article

Implementing Government Elementary Math Exercises Online: Positive Effects Found in RCT under Social Turmoil in Chile

Roberto Araya [1,*] **and Karina Diaz** [1,2,*]

[1] Center for Advanced Research in Education, Institute of Education, Universidad de Chile, Santiago 8330014, Chile
[2] Teachers College, Columbia University, New York, NY 10027, USA
[*] Correspondence: roberto.araya@ciae.uchile.cl (R.A.); kgd2118@tc.columbia.edu (K.D.)

Received: 23 July 2020; Accepted: 4 September 2020; Published: 11 September 2020

Abstract: The impact of online math programs depends on its implementation, especially in vulnerable populations from developing countries. An existing online platform was adapted, at the request of the Chilean Ministry of Education, to exclusively include exercises previously designed and tested by a paper-based government program for elementary school. We carried out a cluster-randomized controlled trial (RCT) with 50 fourth grade classrooms. Treatment classrooms used the platform in a weekly 90-min math session. Due to a social instability outbreak in the country, a large unexpected disruption with huge absenteeism occurred in the second half of the semester, which turned this study into a unique opportunity to explore the robustness of the platform's effects on students' learning. Using multiple imputation and multilevel models, we found a statistically significant effect size of 0.13, which corresponds to two extra months of learning. This effect is meaningful for four reasons. First, it has double the effect of the paper-based version. Second, it was achieved during one semester only. Third, is half that obtained with the platform for a complete year with its own set of exercises and with two sessions per week instead of one. Fourth, it was attained in a semester with a lot of absenteeism.

Keywords: RCT; educational technology; blended learning; testing and evaluation; multilevel imputation

1. Introduction

There is ample evidence to suggest that education has not dramatically changed over recent centuries. Even after the introduction of textbooks, students continue to spend their class time by primarily listening to lectures and taking notes. Why does education seem so immune to transformations? Labaree [1] argues that education is a far more complex domain than other areas. For example, he compares a typical nuclear power facility with a school. Since every component of a nuclear facility is causally interrelated with the others, it is much easier to trace the source of any deficiencies and fix them accordingly. Schools, conversely, are composed of completely independent units: isolated classrooms. If one classroom performs well, it does not immediately produce an effect in parallel classrooms. Superintendents and principals generally track mean performance across classrooms, and, on average, good and bad performances cancel each other out. As a whole, a school therefore remains highly stable.

However, after several decades of experimental studies introducing ICT for math teaching and learning in K–12, there is still a wide range of impacts. For example, a meta-analysis of 71 evaluations in the United States reported effects by time of use [2]. The study shows that for evaluated programs where students spent less than 30 min a week, the average effect was 0.06 SD; where students spent between 30 and 75 min, it was 0.20 SD; finally, where students spent more than 75 min, contrary to what one would expect, the result was 0.14 SD. In a more recent study, [3] reports 14 studies that strongly emphasize the use of technology. Most of them rotate students through technology and non-technology activities. The weighted mean effect size was +0.07. A 2019 study in 26 municipalities in Sweden [4] found no significant impact of an ICT program on standardized tests in mathematics or language on average, but it could, unfortunately, increase inequality in education. Further, a systematic review on 85 independent evaluations found that shorter ICT programs were much more effective in promoting mathematics achievements than longer ones with a mean effect size of 0.35 SD [5]. The theoretical framework for the study of ICT in schools highlights the importance of the implementation process and the context in which this implementation is situated [6]. The integration and final adoption of technological tools relies heavily on these factors. This framework has been supported by empirical evidence of the effect of practice with immediate feedback from peers and teachers and the inclusion of writing justifications for math problems [7,8]. In developing countries, results are also diverse. A review of experimental evaluations in developing countries focused on mathematics [9], reported effect sizes ranging from 0.14 SD for programs with 80 min of weekly computer time in China, to an effect size of 0.35 SD for a program with 120 min of weekly practice in India, and another program with an effect size of 0.28 for 300 min spent using computers during after-school sessions. However, another 300 min per week program in India had an effect size of −0.48. Cristia et al. [10] and Beuermann et al. [11] studied a randomized experiment with a 1:1 program in poor regions of rural Peru and found no significant impact on test scores in mathematics or language. De Melo, Machado, and Miranda [12] found no effects on math or reading scores in the national implementation of a 1:1 program in primary schools in Uruguay. This divergent variety of effects sizes points to the possibility of strong dependence on the type of implementation of the programs.

According to the UNESCO 2013 TERCE assessment, Chile has the highest national average in 6th grade mathematics in Latin America [13]. However, the Programme for International Student Assessment (PISA) test for fifteen years old students, positions Chile in 59th place out of 78 participating countries. Further, its score is not statistically significantly different from scores of countries such as Kazakhstan, Moldova, Baku (Azerbaijan), Thailand, Uruguay, and Qatar [14]. Araya et al. [7], present evidence and theoretical reasons to back the claim that guided technology programs focused on practice, can be effective, efficient, and relatively easy to scale up, under the Chilean context. In [15], the use of the same platform with the originally designed platform exercises was reported. The effect was computed in 15 fourth grade classes from 11 vulnerable Chilean schools where the platform was used during the full educational year. Measured with the National Standardized test, a paper-based assessment implemented by an independent government agency yearly in all schools, the improvement over previous years was 0.26 SD higher than the national improvement.

Later, in Araya et al. [16], reported the results of three years of implementation of the same platform in 11 public schools from a low SES urban district in Chile. This included 43 fourth grade classes and 1355 students. Improvement over previous years on the National Standardized fourth grade math test was 0.28 SD higher than the improvement made by a neighboring district with a similar population. Next, in [17], eight years of use of the same platform and exercises were analyzed. The authors found that on the national standardized test scores, the 80 classes that were under treatment obtained 0.30 SD higher results than on the 32 classes that were not treated. In a more recent study, Araya et al. [18], experimentally evaluated the platform with a RCT with 48 classes from 24 low-performing primary schools in Chile, where at each school a class was randomly assigned to treatment. It was implemented with two weekly sessions in a computer lab during the whole educational year. The impact was measured with the Chilean National Standardized Exam by using a multilevel model, and a positive effect on math learning of 0.27 SD was found.

Moreover, the Ministry of Education in 2011 and 2012 implemented a paper-based implementation of the "Plan de Apoyo Compartido" (PAC) program. This is a standardized teaching material program that included the support of internal and external pedagogic teams. It was implemented in under-performing schools in Chile. In [19], Bassi et al. conducted a RCT to estimate the effectiveness of PAC. The intervention improved performance in math for the first cohort of students (effect size of 0.068 and it was statistically significant), but not in the second cohort. Thus, in this paper we study the impact of exclusively using PAC exercises in the ConectaIdeas platform, instead of the standard ConectaIdeas exercises that were previously designed and tested. According to Bowen [20], the need for customizable platforms that allow teachers to customize materials is perhaps the largest obstacle to widespread adoption of interactive online learning. This study can help determine whether the effect is due to the exercises or the implementation in an online platform.

Hill et al. [21] reviewed experimental evaluations in education in the US and documented that the average effect on broad standardized tests was 0.07 SD, compared to an average effect of 0.23 SD for narrow standardized tests and to 0.44 SD for specialized tests developed for specific interventions. According to Cheung et al. [22] effect sizes are roughly twice as large for published articles, small-scale trials, and experimenter-made measures than for unpublished documents, large-scale studies, and independent measures, respectively. In addition, effect sizes are significantly higher in quasi-experiments than in randomized experiments. Moreover, across seven WWC-accepted math studies, the mean effect size was +0.45 for measures with treatment-inherent measures and −0.03 for measures used in the same studies that were not inherent to the treatment [22].

In this paper, we explore the use of an online platform in an unforeseen environment. First, instead of using the originally designed and improved exercises for the platform, this study implements paper-based exercises designed by the Chilean Ministry of Education. These are valuable exercises product of an extensive recompilation, updated and upgraded in a previous program developed by the Ministry of Education. Moreover, this upgraded program and its exercises have been previously studied [19] and have shown positive results in math during its first year of implementation, but not for the second year. According to [19], a possible explanation for this decline is the decrease in rigor of implementation compared to the first year. The first research question is then to estimate whether the effect size is maintained or hopefully increased in the online version. Second, in the middle of the semester and until the end of the implementation, a huge social outbreak shook the country. Several schools closed due to teacher strikes and to social unrest. Thus, the second research question aims to determine whether the online version platform could still impact student learning under this unstable condition, which involved a huge level of student absenteeism, and how much erosion occurred when comparing with effects of previous evaluation of the use of the same online platform in math, for fourth graders.

Particularly in this study, we used a large-scale test to estimate the impact of the program on students' outcomes. The main contribution of this paper is to measure the effect of a platform when a completely new set of exercises are exclusively used, or from another point of view, when a set of materials with exercises is used in an online platform. Moreover, the social turmoil during the second half of the implementation period had a huge impact on attendance, which turned this study into a rare opportunity to estimate the robustness of the effect of the intervention under difficult contextual circumstances. Missing data was one of the main unexpected challenges faced because of the social turmoil. This paper illustrates the application of multilevel multiple imputation models to deal with missingness in the outcome variable together with the use of multilevel regression models to estimate the program effect size. Finally, we analyzed the effect of the inclusion of at least one open-ended question in each session with written answers and peer review.

2. Methods

2.1. Sample and Implementation

According to [23], a minimum of 30 to 60 teachers or schools is necessary in order to have sufficient statistical power to detect at least medium-sized effects and allow for attrition. In this study, we purposively recruited 50 fourth grade classrooms from undeserved schools of several Santiago districts in July 2019; 77 percent of the participating schools are publicly funded, and 23 percent are voucher schools. Selection criteria included: time and space allocations for the use of ICT for math teaching; school administrators who are open to the implementation of ICT; willingness of teachers to engage in the use of ICT and to classroom visitations; and school technological infrastructure, including an internet connection. Half of the schools were randomly assigned to the treatment. A total of 659 students participated in the treatment and 538 were part of the control group. The average age of the students who participated ranges from 9 to 10 years old. Participating teachers had no previous experience with the ConectaIdeas platform. Teachers in the treatment group were assigned to an initial training where they were introduced to the platform and final objectives of the project. The implementation team secured official written consent for participation, classroom observation, and data collection from administrators and legal guardians of participating students.

Given the short time available for the pilot implementation (from August to November), the random selection was performed before results from the pretest were available. The results of this test took almost a month after the last classroom completed the test in mid-August. To allow sufficient implementation time, the program started right after the baseline was measured. Thus, the treatment groups balance was later verified. As illustrated in Table 1, the control group had slightly lower results, which corresponds to 0.04 standard deviations in the pretest. However, these differences were not statistically significant (t(1195) = 0.783, p = 0.433).

Table 1. Treatment groups baseline.

Group	N	Pretest Mean	SD	t	df	*p*-Value
Treatment	659	560.55	43.59	0.783	1195	0.433
Control	538	558.59	42.84			
Total	1197	559.67	43.25			

Previous studies show that using technology that allows for immediate feedback in the classroom can have a positive impact on students' outcomes, particularly in math and science [24,25]. The ConectaIdeas platform was designed to drive the progression of the classroom as a whole, and not to leave students by themselves. It provides a real time early alert system that lists students who are having more difficulties during the session and promotes peers' cooperation. Thus, using the automatic early alert system, teachers can assist those students or "assign" them to students who are ahead of their peers (i.e., students that finish early and perform well) for help. Students being assisted by peers can in turn evaluate the quality of the support. This system allows teachers and lab coordinators to work with both types of students in order to improve their understanding of the content, as well as their communication skills.

The platform also detects if there are exercises that are proving difficult for the whole class. Permitting teachers and lab coordinators to freeze the system and explain the necessary concepts. All exercises are related to specific Learning Objectives of the National Curriculum. Particularly in this implementation, all the exercises of the PAC program were assigned to those specific Learning Objectives. After each session, teachers receive a report of the Learning Objectives coverage. A particular feature of the implementation on this platform is to promote reflection and written argumentation, as well as peer review of those written arguments. In each session, teachers ask at least one open question. Students answer on their devices, and then they have to randomly review the answers of one of their classmates. Implementation started in mid-August and lasted until the first week of December, but during 2 to 4 weeks in late October the average length was zero, given that classes were interrupted due to the social turmoil. As shown in Figure 1, the average length of students' answers to math problems ranged between 9 to 16 words and excluding December, where there was just one week of implementation, there was a positive growth trend for this average. Moreover, the difference between the mean answer length in August and November was statistically significant (t(892) = 4.184, p < 0.001).

Class attendance of each student has a direct effect on students' learning, but it can also affect classmates. According to Gottfried [26], chronic absenteeism (missing more than 10% of school days) has a damaging effect on students and a potential negative spillover that reduces outcomes for other students in the same classroom. This negative spillover effect responds to the paradigm of teachers' time and instruction being a public good [27] i.e., something that is 'consumed' by all students in the classroom. Thus, greater chronic absenteeism produces a big disruption in instruction, and then it consumes the efforts and time of teachers attending to those students when they return.

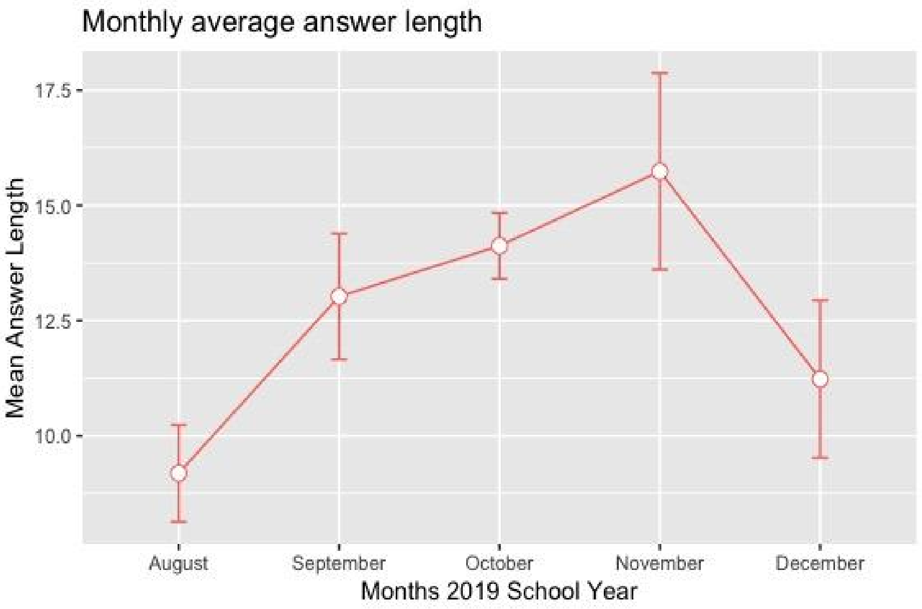

Figure 1. Math problems average answer length by month with 95% confidence intervals.

The implementation was led by an experienced teacher that supervised two new lab coordinators. In each treatment classroom, a lab coordinator supported or commanded the sessions. Both lab coordinators were elementary school teachers who had never worked on the platform before, and who were trained on the job during the sessions of the first week. Each lab coordinator visited 12 or 13 classrooms each week and provided on the job training to participating teachers. The treatment session lasted 90 min, and it was completed in one of the weekly regular math sessions. Thus, there was no increase in instructional time. The overall mean number of math exercises completed by students was 364, September being the month with the highest average number (105). The reduction in the number of math problems during the following months—October and November—can be explained by the social turmoil and its impact in student attendance. In fact, the difference between the mean number of math problems completed by students in September was significantly higher than in October (t(1252) = 5.7, $p < 0.000$) and November (t(1145) = 9.9, $p < 0.000$). As previously discussed, during December, the program was only implemented for one week, which explains the lower number of exercises, as can be seen in Figure 2.

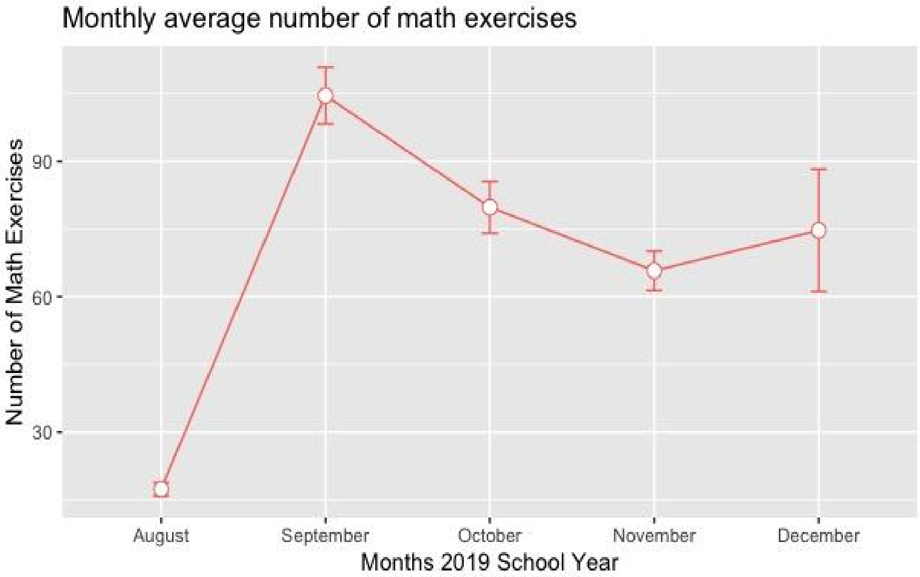

Figure 2. Monthly average number of completed math exercises with 95% confidence intervals.

Students attendance was measured before and during the implementation. As shown in Figure 3, both treatment and control groups followed a similar pattern of mean monthly attendance. The average attendance was lower during the second academic semester (August-December) for both groups, with a mean difference of four missing days per month. The reduction between first and second semester was significant for the treatment (t(1957) = 17.1, p < 0.000) as well as the control group (t(847) = 12.2, p < 0.000) and the treatment group had, on average, a lower attendance level than the control group in the second semester. Moreover, 11 percent of the students in a classroom missed more than 10 percent of school days during the second semester, on average. Following Gottfried [26] guidelines, we anticipate this having a negative impact on students' outcomes.

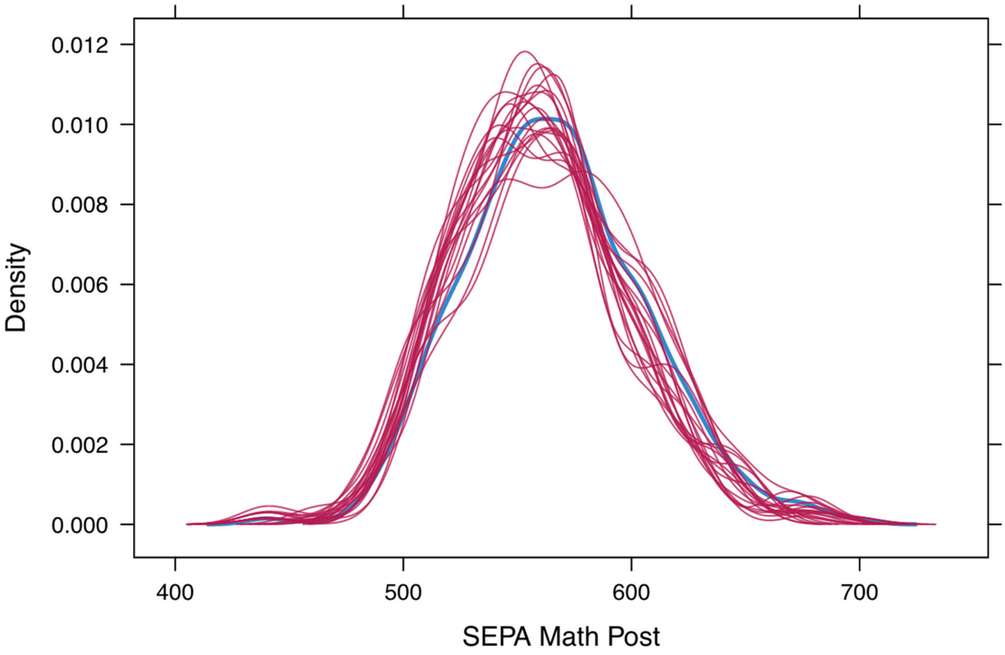

Figure 3. Multiple imputed SEPA-posttest distribution vs. observed values.

Finally, the implementation was carried out meeting the following criteria [3]:

1. Students who qualified for special education services but attended mainstream mathematics classes were included.
2. Random assignment to treatment and control.
3. Control groups used an alternative program already in place, or "business-as-usual".
4. The treatment program was delivered by ordinary teachers, not by the program developers, researchers, or their graduate students.
5. Pretest differences between experimental and control groups were less than 25% of a standard deviation. Indeed, the difference was just 4% of a standard deviation.
6. Differential attrition between experimental and control groups from pre-post-test was 10%, which is less than the limit of 15% suggested [3].
7. Assessments were not made by developers of the program or researchers. They were designed and administered by a regular provider of the Ministry of Education, with the most experience in the country, and who also is a provider of tests of the UNESCO ERCE 2019 [28] test for Latin America.
8. The study had more than two teachers and 30 students in each condition. Indeed, there were 18 teachers in the Treatment Group, another 18 teachers in the Control Group, and a total of 1197 students.
9. The study had more than 12 weeks of duration.
10. Additionally, the intervention in the treatment group was in regular class hours, not in extra supplementary time.

2.2. Analysis

A third party, an Item Response Theory (IRT) calibrated SEPA test, was used as our pre and post outcome measure. SEPA was developed by MIDE UC, Universidad Católica de Chile, and allows measuring the progress of student learning throughout the school year across a set of tests, based on the Chilean curricular framework from 1st to 11th grade in Language and Mathematics. SEPA defines a Reference Sample (RS) to achieve a better representativeness of the national distribution of schools according to dependency, socioeconomic level, and school performance (in the National Standardized Test, SIMCE), and then fits an IRT model using the RS to estimate a standard score for each student. The fourth grade SEPA test has been validated and has a reliability of 0.91 [29].

Given the nested structure of the data, a three-level Hierarchical Linear Model (HLM) was specified to explore the effect size of the intervention and student's academic achievement. HLM is commonly used in education research, mainly because of the need to take aggregation levels into account in order to comprehend the differences observed between students [30] and thus allowing an unbiased significance test [31]. Following HLM procedures, first an unconditional model (Equation (1)) was estimated to clarify whether HLM is appropriate for the data:

$$\begin{gathered} Y_{ij} = \gamma_0 + \mu_j + \epsilon_{ij} \\ \mu_j \sim N\left(0, \sigma_\mu^2\right) and\ \epsilon_{ij} \sim N\left(0, \sigma_\epsilon^2\right),\ all\ independent, \end{gathered} \tag{1}$$

where Y_{ij} is the ith observation in the jth group (class/school level) estimated student outcome, γ_0 is the unobserved overall mean, μ_j is the unobserved random effect shared by all values in group j, and ϵ_{ij} is the student-level residual term.

The intraclass correlations (ICC) is 0.149, in other words, 15% of the total variance of the SEPA math test is explained by classroom differences. Further, 2.8% of the total variance is explained by school differences. Based on these results and following the guidelines of considering clusters with ICC as low as 0.01, we included schools and classrooms clusters in our three-level model [32].

Second, a full model was specified to examine the effect of factors at the student, classroom, and school-level (Equation (2)):

$$\begin{gathered} Y_{ijk} = \gamma_{0jk} \; + \beta_{1jk} STUDENT_{ijk} + \beta_{njk} STUDENT_{ijk} + \epsilon_{ijk} \\ \gamma_{0jk} = \gamma_{ook} + \mu_{0jk} \\ \gamma_{00k} = \pi_{000} + r_{00k} \\ \mu_{0jk} \; \sim N\left(0, \sigma_{\mu}^{2}\right) and \; \epsilon_{ijk} \; \sim N\left(0, \sigma_{\epsilon}^{2}\right), \; all \; independent, \end{gathered} \tag{2}$$

where Y_{ijk} is the ith observation in the jth classroom at the kth school, $\beta_{1jk} \ldots \beta_{njk}$ refers to the fixed effect (slope) of the student level variables $\left(STUDENT_{ijk}\right)$, γ_{00k} refers to the class-level random intercept (i.e., grand mean of scores), and π_{000} refers to the school-level random intercept.

We followed the methods used by the National Center for Education Evaluation (NCEE) Technical Methods report [33] in order to address the problem of missing data in the analysis of data in Randomized Controlled Trials (RCTs) of educational interventions, with a particular focus on the common educational situation in which groups of students such as entire classrooms or schools are randomized. Table 2 describes the proportion of missing cases in each treatment group.

Table 2. Post-test missingness proportion by treatment group.

	Treatment Group		**Control Group**		**Total**	
Measure	n	% Missing	n	% Missing	n	% Missing
SEPA-Post	659	15%	538	25%	1197	20%

For the purpose of this analysis, we made the assumption that the missing data follows a Missing at Random mechanism (MAR) [33], meaning the probability of being missing is the same only within groups defined by the observed data. This would imply that there is a systematic relationship between the propensity of missing values and the observed data [34]. Further, once one has conditioned on all the observed data, any remaining missingness is completely random. Consequently, when the cause of missingness is taken into account, MAR missingness leads to unbiased parameter estimates [35]. Table 3 presents a comparison between observed values of students who completed the SEPA-posttest and those who did not. We used the Kruskal Wallis test for the continuous variables and the Chi-squared test for categorical variables to determine significant differences between both groups. Results suggested that there is a significant relationship between observed variables and missingness in SEPA-post outcomes, which sustains our claim of the data not being MCAR and thus a MAR assumption being more suitable [36,37].

Table 3. Comparison between values of responders (not missing) and non-responders (missing) on the SEPA post-test.

Measure		**Not Missing**	**Missing**	***p***
SEPA Math Pre	Mean (SD)	561.4 (43.5)	552.6 (41.7)	0.005
Group	Control	401 (41.8)	137 (57.6)	<0.001
	Treatment	558 (58.2)	101 (42.4)	
Sex	Female	495 (51.6)	116 (48.7)	0.47
	Male	464 (48.4)	122 (51.3)	
GPA	Mean (SD)	5.9 (0.5)	5.8 (0.7)	0.002
Attendance	Mean (SD)	90.7 (7.2)	84.1 (12.1)	<0.001
Completed Exercises	Mean (SD)	215.6 (279.5)	144.8 (230.5)	<0.001
Answer Length	Mean (SD)	7.6 (10.0)	5.4 (11.2)	0.003

Complete case deletion is probably the most commonly used procedure when dealing with missing data. However, when the missingness mechanism fails to meet the Missing Completely at Random assumption, as in our case, complete case deletion will yield bias estimates [34,38]. Further, the loss of sample members can reduce the power to detect statistically significant differences. Instead, we used the multiple imputation approach, which has become the method of choice in many contexts of missing data [39].

Ignoring the clustering and imputing the data by a one-level approach will underestimate the ICC [40–42] and that, in certain cases, can be more harmful than complete case deletion, due to wrong model specification [43]. Thus, we used the multilevel predictive mean matching method, which uses linear mixed models with random draws from the regression coefficients and the random effects to impute missing outcomes [44,45]. Moreover, following the literature recommendation, we included all the available complete variables in the data set in order to capture the assumption of MAR [46,47]. Based on previous recommendations [48,49], we first generated 20 imputations. Later, we analyzed each completed data set using the HLM model previously detailed. Finally, using Rubin's rules implemented in the 'mice' R package, we combined the estimates from the analyses and obtained our effect size estimates [50].

3. Results

Relevant student-level covariates in this study include continuous variables such as SEPA-math baseline (SEPA Math Pre), overall attendance, grade point average (GPA), the total number of performed math exercises (NumberExercises), and average length of open-ended math questions (AnswerLength), as well as sex and treatment group indicators. Descriptive statistics for each predictor are presented in Table 4. On average, students scored 559.67 points in the pretest and had a mean GPA of 5.87. In general, mean attendance was 89.42 percent and the average number of platform exercises was 202. Open-ended questions had a length of 7.2 words on average. Further, the correlation between pre-test and post-test scores was 0.72.

Table 4. Relevant covariates descriptive statistics.

Measure		
SEPA Math Pre	Mean (SD)	559.67 (43.3)
Group	Control	538
	Treatment	659
Sex	Female	611
	Male	586
GPA	Mean (SD)	5.87 (0.6)
Attendance	Mean (SD)	89.42 (8.8)
Number Exercises	Mean (SD)	201.52 (271.8)
Answer Length	Mean (SD)	7.19 (10.26)

Equation (3) specifies the linear mixed model used to impute missing SEPA-math post results. Further, Figure 3 illustrates the distribution of SEPA-math post scores for observed (blue) and imputed (red) values after generating twenty multiple imputed datasets. Results suggest that the imputed SEPA-math post imputed values follow a distribution similar to that shown in the observed values:

$$\begin{gathered} Y_{ij} = \gamma_{0j} \ + \beta_{1j} SEPAMathPre_{ij} \ + \beta_{2j} Group : Treatment_{ij} \\ + \beta_{3j} Sex : Male_{ij} + \beta_{3j} GPA_{ij} + \beta_{4j} Attendace_{ij} + \\ \beta_{5j} NumberExercises_{ij} + \beta_{6j} AnswerLength_{ij} + \epsilon_{ij} \\ \gamma_{0j} = \gamma_{oo} + \mu_j \\ \mu_j \ \sim N\left(0, \sigma^2_\mu\right) and \ \epsilon_{ij} \ \sim N\left(0, \sigma^2_\epsilon\right), \ all\ independent. \end{gathered} \tag{3}$$

For each completed data set, the implementation effect size was estimated by fitting the HLM shown in Equation (4). Later, the estimates from each analysis were combined following Rubin's rules. Results showed a positive significant effect of the treatment on SEPA-math post scores (t(78) = 2.802, *p* = 0.035). The intervention effect size was estimated using the covariate adjusted mean difference (regression coefficient) and the unadjusted post-test standard deviation. Thus, the estimated treatment effect size was 0.13 SD and had a variance of 0.0016. Moreover, SEPA pre-test results (t(79) = 16.49, p = 0.000) and overall GPA (t(35) = 3.85, p = 0.000) were also significant and had a positive effect on students post-test scores. On the other hand, male students on average showed 2.6 points higher than female students for these scores, but this difference is not significant (t(186) = 1.34, p = 0.182). Attendance appears to have had a negative effect on overall results (t(136) = −0.41, p = 0.002). Table 5 summarizes the final effect size estimates from the HLM model:

$$\begin{aligned} Y_{ijk} &= \gamma_{0jk} + \beta_{1jk}PreTest_{ijk} + \beta_{2jk}Group : Treatment_{ijk} \\ &+\beta_{3jk}Sex : Male_{ijk} + \beta_{3jk}GPA_{ijk} + \beta_{4jk}Attendace_{ijk} + \epsilon_{ijk} \\ \gamma_{0jk} &= \gamma_{ook} + \mu_{0jk} \\ \gamma_{00k} &= \pi_{000} + r_{00k} \\ \mu_{0jk} &\sim N\left(0, \sigma^2_\mu\right) \mathit{and}\ \epsilon_{ijk} \sim N\left(0, \sigma^2_\epsilon\right),\ \mathit{all\ independent}. \end{aligned} \tag{4}$$

Table 5. SEPA-posttest HLM model 4 results.

	Estimate	**Std. Error**	**t**	***p***
Intercept	210.055	16.842	12.472	0.000
SEPA-Math PRE	0.534	0.032	16.491	0.000
Group: Treatment	5.615	3.470	4.618	0.019
Sex: Male	2.689	2.005	1.341	0.182
GPA	14.864	3.856	3.855	0.000
Attendance	−0.405	0.131	−3.099	0.002

Similarly, we estimated the effect size of the average length of students' answers to math problems by fitting the HLM presented in Equation (5) for each completed data set and then pooling the results. Findings suggested a positive and significant effect of the answer length on SEPA-math post scores (t(222) = 2.053, p = 0.041). Moreover, the effects of SEPA pre-test results, overall GPA, attendance, and sex followed the same patterns as in Model 4 estimates. Table 6 summarizes these results.

$$\begin{aligned} Y_{ijk} &= \gamma_{0jk} + \beta_{1jk}PreTest_{ijk} + \beta_{2jk}AnswerLength_{ijk} \\ &+\beta_{3jk}Sex : Male_{ijk} + \beta_{3jk}GPA_{ijk} + \beta_{4jk}Attendace_{ijk} + \epsilon_{ijk} \\ \gamma_{0jk} &= \gamma_{ook} + \mu_{0jk} \\ \gamma_{00k} &= \pi_{000} + r_{00k} \\ \mu_{0jk} &\sim N\left(0, \sigma^2_\mu\right) \mathit{and}\ \epsilon_{ijk} \sim N\left(0, \sigma^2_\epsilon\right),\ \mathit{all\ independent}. \end{aligned} \tag{5}$$

Table 6. SEPA-posttest HLM model 5 results.

	Estimate	**Std. Error**	**t**	***p***
Intercept	214.632	17.211	12.471	0.000
SEPA-Math PRE	0.531	0.033	15.99	0.000
Answer Length	0.225	0.109	2.053	0.041
Sex: Male	2.864	2.037	1.406	0.161
GPA	14.661	3.850	3.808	0.001
Attendance	−0.406	0.131	−3.101	0.002

4. Discussion

The aim of this study was to estimate whether the impact of the use of nationally developed math exercises is maintained or even increased when integrated in an online platform. Results show a 0.13 SD positive effect size of the implementation on students' outcomes when measured with a large-scale test SEPA-math. This effect corresponds to almost two extra months of learning after translating learning gains according to US year-long learning gains in math for fourth graders [51]. Further, the effect size achieved by the online platform intervention was double the effect achieved by using the same exercises on a paper version for a whole year.

The vast majority of Chilean urban schools have fiber optic internet connection, which allowed us to convert a paper-based government program of math exercises to an online version. Further, the selection criteria applied in this project, made it possible to ensure that all participating schools had the technological infrastructure required for the use of the online platform in the classroom. During the implementation, students were able to carry out all the activities without internet connection problems.

According to Gottfried [26], chronic absenteeism can not only have a negative effect on students missing excessive school days, but also has the potential to lower outcomes for other students within a similar educational context. The results shown in this study shed light on the effectiveness of the ContectaIdeas platform under unstable conditions due to the social outbreak, despite the increase in absenteeism during the application period. Moreover, the estimated effect size is almost half the impact achieved with two sessions per week for a complete year when using the same online platform but using pre-designed exercises. Thus, the current implementation has shown to be promising when compared with effects of previous evaluations on the use of the same online tool for fourth graders.

Further, as discussed by Kuhfeld et al. [52], the effect size in the second semester is 0.01 SD lower than the effect size in the first semester in 4th grade in the US, even though the second semester is longer. Thus, the Average Monthly RIT Gains is 2.00 in the second semester and 2.02 in the first semester. We can then estimate that the yearly overall effect size for this implementation would have been 0.27 SD. However, this estimation does not consider the extra absenteeism in the second semester due to the social outbreak, we could then estimate that under normal conditions, the yearly overall impact of the implementation would have been even higher.

There is evidence that writing can improve learning. A meta study with 6th to 8th graders of 48 writing-to-learn programs [53] shows that writing can have a small, positive impact on conventional measures of academic achievement. According to the authors, writing can prompt and support the use of rehearsal strategies, elaboration strategies, organization strategies, and comprehension-monitoring strategies. In another more recent meta-analysis of 12 studies Bicer et al. [8] found an overall effect size of 0.42. Similarly, our findings show a significant positive effect of average length of students' answers to math problems on math learning. Likewise, recent studies show that incorporating real time monitoring and feedback into online platforms can have a positive impact on overall students' outcomes in math [24,25]. We argue that both components are an essential part of the positive results of the ConectaIdeas online implementation presented here.

Finally, this paper provides evidence of the positive impact of incorporating regular paper-based math exercises into an online platform, as well of the robustness of the effect of an intervention under unique contextual circumstances. Furthermore, it exemplifies the use of multilevel multiple imputation models to handle missing data in the outcome variable, as opposed to purposely deleting observations—complete case deletion—which would have reduced the power of our study and biased the estimated results.

5. Conclusions

The implementation evaluated in this work has important practical implications. First, converting paper-based mathematic exercises—previously used and refined for years by the Ministry of Education—to an online platform, proved to improve the effectiveness of such exercises. This effect is doubled and its significant despite the fact that the number of sessions was reduced from twice to once per week, and the fact that the intervention only lasted one semester. Moreover, the effect was achieved despite the social turmoil that affected the country in the middle of the semester that increased absenteeism to levels much higher than the historical ones.

Second, in each session students were required to answer at least one open question, which included arguments of the procedures and the logic used to solve the problem. These written answers were shared with their peers, who reviewed and commented on the answers. This activity shown to have an effect on student learning. These results contribute to informing policy decisions regarding the use of existing math exercises under an online platform.

Although these findings have shown promise, there are several aspects that require further study and will be addressed in future work. For instance, studying not only the length of the written answers but how they relate to the type of question posted by the teacher. Araya R., et al. [54] addressed this issue and found that the presence of certain keywords in the question demonstrated to be relevant. However, it is necessary to further extend the study of type of questions using topic models or the natural language processing methods. It is also necessary to analyze the type and "quality" of answers given by students and its relationship to learning.

A second aspect that needs to be further studied is the effect of the strategy of peer collaboration through student assistants, implemented in ConectaIdeas. In each session, a platform module preselects students who are performing well to become candidates for classroom assistants. A couple of students are then selected by the teacher to be teaching assistants during the session. Students can then request help from any assistant or the teacher itself to solve an exercise. Once the assistant is finished, students can evaluate the quality of the help received, and the teacher assistant can also evaluate how well he or she thinks the person who helped understood the explanation. Evaluating the impact of this strategy will require a different experimental study.

A third feature that is important to address is the impact of the platform on teachers' didactic strategy. In Araya R., et al. and Uribe P., et al. [55–57], various classroom observation protocols are used to classify each moment of the session, and different machine learning algorithms are also used to perform automatic analysis of teaching discourse transcriptions. We have been using both methodologies to determine the impact of the use of platforms on teaching strategies. This is work in progress.

Finally, one of the main limitations of this implementation is related to its sustainability and was revealed this year during the quarantine in response to COVID-19. Although most urban underserved schools in Chile have optic fiber internet connections, students at home have very unstable internet. In addition, a big proportion of them rely on their parents' smartphones for internet connections. Even though the ConectaIdeas platform requires very little internet bandwidth, it does need a stable connection. Thus, the challenge is to adapt the platform to work offline and to accommodate both the interface and the exercises, to facilitate its use on small screen devices. In a future study, we will analyze an offline version of the platform for smartphones that is now being tested by students from vulnerable sectors in Chile and Peru.

Author Contributions: Conceptualization, Methodology, Analyses, Writing: R.A. and K.D.; Funding acquisition, R.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the UCE, Ministry of Education of Chile and the Chilean National Agency for Research and Development (ANID).

Acknowledgments: Support from ANID/PIA/Basal Funds for Centers of Excellence FB0003 is gratefully acknowledged.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in writing of the manuscript, or in the decision to publish the results. R.A. was part of the ConectaIdeas platform design team.

References

1. Labaree, D. *Someone Has to Fail: The Zero-Sum Game of Public Schooling*; Harvard University Press: Cambridge, MA, USA, 2010.
2. Cheung, A.C.K.; Slavin, R.E. The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educ. Res. Rev.* **2013**, *9*, 88–113. [CrossRef]

3. Pellegrini, M.; Lake, C.; Inns, A.; Slavin, R. Effective Programs in Elementary Mathematics: A Best-Evidence Synthesis. Available online: http://www.bestevidence.org/word/elem_math_Oct_8_2018.pdf (accessed on 23 June 2020).
4. Hall, C.; Lundin, M.; Sibbmark, K. A Laptop for Every Child? The Impact of ICT on Educational Outcomes. Available online: https://www.ifau.se/globalassets/pdf/se/2019/wp-2019-26-a-laptop-for-every-child-the-impact-of-ict-on-educational-outcomes.pdf (accessed on 23 June 2020).
5. Li, Q.; Ma, X. A Meta-analysis of the Effects of Computer Technology on School Students' Mathematics Learning. *Educ. Psychol. Rev.* **2010**, *22*, 215–243. [CrossRef]
6. Lim, C.P. A theoretical framework for the study of ICT in schools: A proposal. *Br. J. Educ. Technol.* **2002**, *33*, 411–421. [CrossRef]
7. Araya, R.; Cristia, J. Guiding Technology to Promote Student Practice. In *Learning Mathematics in the 21st Century: Adding Technology to the Equation*; Arias Ortiz, E., Cristia, J., Cueto, S., Eds.; Inter-American Development Bank: Washington, DC, USA, 2020; pp. 225–253.
8. Bicer, A.; Perihan, C.; Lee, Y. The Impact of writing practices on students' mathematical attainment. *Int. Electron. J. Math. Educ.* **2018**, *13*, 305–313. [CrossRef]
9. Arias Ortiz, E.; Cristia, J. The IDB and Technology in Education: How to Promote Effective Programs? Available online: https://publications.iadb.org/publications/english/document/The-IDB-and-Technology-in-Education-How-to-Promote-Effective-Programs.pdf (accessed on 23 June 2020).
10. Cristia, J.P.; Ibarraran, P.; Cueto, S.; Santiago, A.; Severin, E. Technology and Child Development: Evidence from the One Laptop per Child Program. Available online: https://publications.iadb.org/publications/english/document/Technology-and-Child-Development-Evidence-from-the-One-Laptop-per-Child-Program.pdf (accessed on 23 June 2020).
11. Beuermann, D.W.; Cristia, J.P.; Cueto, S.; Malamud, O.; Cruz-Aguayo, Y. One laptop per child at home: Short-term impacts from a randomized experiment in Peru. *Am. Econ. J. Appl. Econ.* **2015**, *7*, 53–80. [CrossRef]
12. De Melo, G.; Machado, A.; Miranda, A. *The Impact of a One Laptop per Child Program on Learning: Evidence from Uruguay*; IZA Discussion Paper 8489; Institute of Labor Economics (IZA): Bonn, Germany, 2014. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2505351 (accessed on 23 June 2020).
13. Valverde, G.; Marshall, J.; Sorto, A. Mathematics Learning in Latin America and the Caribbean. In *What Are the Main Challenges for Mathematics Learning in LAC?* Arias Ortiz, E., Cristia, J., Cueto, S., Eds.; Inter-American Development Bank: Washington, DC, USA, forthcoming.
14. OECD. *PISA 2018 Results (Volume I): What Students Know and Can Do*; OECD Publishing: Paris, France, 2019. [CrossRef]
15. Araya, R.; Van der Molen, J. Impact of a blended ICT adoption model on Chilean vulnerable schools correlates with amount of online practice. In Proceedings of the Workshops at the 16th International Conference on Artificial Intelligence in Education AIED 2013, Memphis, TN, USA, 9–13 July 2013.
16. Araya, R.; Gormaz, R.; Bahamondez, M.; Aguirre, C.; Calfucura, P.; Jaure, P.; Laborda, C. ICT supported learning raises math achievement in low socioeconomic status schools. In *Design for Teaching and Learning in a Networked World*; Conole, G., Klobučar, T., Rensing, C., Konert, J., Lavoué, E., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2015; Volume 9307, pp. 383–388. [CrossRef]
17. Araya, R. Teacher Training, Mentoring or Performance Support Systems. In *AHFE 2018: Advances in Human Factors in Training, Education, and Learning Sciences*; Nazir, S., Teperi, A.M., Polak-Sopińska, A., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2019; Volume 785, pp. 306–315. [CrossRef]
18. Araya, R.; Arias Ortiz, E.; Bottan, N.; Cristia, J. Does Gamification in Education Work? Experimental Evidence from Chile. Available online: https://publications.iadb.org/publications/english/document/Does_Gamification_in_Education_Work_Experimental_Evidence_from_Chile_en_en.pdf (accessed on 23 June 2020).
19. Bassi, M.; Meghir, C.; Reynoso, A. *Education Quality and Teaching Practices*; NBER Working Paper 22719; National Bureau of Economic Research, Inc.: Cambridge, MA, USA, 2016.
20. Bowen, W. *Higher Education in the Digital Age*; Princeton University Press: Princeton, NJ, USA, 2013.

21. Hill, C.; Bloom, H.; Black, A.R.; Lipsey, M.W. Empirical benchmarks for interpreting effect sizes in research. *Child Dev. Perspect.* **2008**, *2*, 172–177. [CrossRef]
22. Cheung, A.C.K.; Slavin, R.E. How methodological features affect effect sizes in education. *Educ. Res.* **2016**, *45*, 283–292. [CrossRef]
23. Roschelle, J.; Feng, M.; Gallagher, H.; Murphy, R.; Harris, C.; Kamdar, D.; Trinidad, G. *Recruiting Participants for Large-Scale Random Assignment Experiments in School Settings*; SRI International: Menlo Park, CA, USA, 2014.
24. Curto Prieto, M.; Orcos Palma, L.; Blázquez Tobías, P.; León, F. Student assessment of the use of kahoot in the learning process of science and mathematics. *Educ. Sci.* **2019**, *9*, 55. [CrossRef]
25. Zhao, X.; van den Heuvel-Panhuizen, M.; Veldhuis, M. Insights chinese primary mathematics teachers gained into their students' learning from using classroom assessment techniques. *Educ. Sci.* **2019**, *9*, 150. [CrossRef]
26. Gottfried, M. Chronic absenteeism in the classroom context: Effects on achievement. *Urban Educ.* **2019**, *54*, 3–34. [CrossRef]
27. Lazear, E. Educational production. *Q. J. Econ.* **2001**, *116*, 777–803. [CrossRef]
28. Estudio ERCE 2019. Evaluación de la Calidad de la Educación en América Latina. Available online: https://es.unesco.org/fieldoffice/santiago/llece/ERCE2019 (accessed on 23 June 2020).
29. Manzi, J.; García, M.R.; Taut, S. *Validez de Evaluaciones Educacionales de Chile y Latinoamérica*; Ediciones UC: Santiago, Chile, 2019.
30. Goldstein, H. Multilevel Modelling of Educational Data. In *Methodology and Epistemology of Multilevel Analysis*; Courgeau, D., Ed.; Methods Series; Springer Science+Business Media: Dordrecht, The Netherlands, 2003; Volume 2, pp. 25–42.
31. Raudenbush, S.W.; Bryk, A. *Hierarchical Linear Models: Applications and Data Analysis Methods*; SAGE Publications, Inc.: Thousand Oaks, CA, USA, 2002.
32. Kreft, I.; de Leeuw, J. *Introducing Statistical Methods: Introducing Multilevel Modeling*; SAGE Publications, Ltd.: London, UK, 1998.
33. Puma, M.; Olsen, R.; Bell, S.; Price, C. *What to Do When Data Are Missing in Group Randomized Controlled Trials*; National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education: Washington, DC, USA, 2009.
34. Schafer, J.L.; Graham, J.W. Missing data: Our view of the state of the art. *Psychol. Methods* **2002**, *7*, 147–177. [CrossRef] [PubMed]
35. Graham, J.W. Missing data analysis: Making it work in the real world. *Annu. Rev. Psychol.* **2009**, *60*, 549–576. [CrossRef] [PubMed]
36. Little, R.J.A. A Test of missing completely at random for multivariate data with missing values. *J. Am. Stat. Assoc.* **1988**, *83*, 1198–1202. [CrossRef]
37. Heymans, M.W.; Eekhout, I. *Applied Missing Data Analysis with SPSS and (R)Studio*; Heymans and Eekhout: Amsterdam, The Netherlands, 2019. Available online: https://bookdown.org/mwheymans/bookmi/ (accessed on 23 June 2020).
38. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*, 2nd ed.; Wiley: New York, NY, USA, 2002.
39. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley & Sons: New York, NY, USA, 1987.
40. Taljaard, M.; Donner, A.; Klar, N. Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biom. J.* **2008**, *50*, 329–345. [CrossRef]
41. Van Buuren, S. Multiple Imputation of Multilevel Data. In *Handbook of Advanced Multilevel Analysis*; Hox, J.J., Roberts, J.K., Eds.; Routledge: Milton Park, UK, 2011; pp. 173–196.
42. Enders, C.K.; Mistler, S.A.; Keller, B.T. Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychol. Methods* **2016**, *21*, 222–240. [CrossRef]
43. Grund, S.; Lüdtke, O.; Robitzsch, A. Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organ. Res. Methods* **2018**, *21*, 111–149. [CrossRef]
44. Vink, G.; Lazendic, G.; van Buuren, S. Partitioned predictive mean matching as a multilevel imputation technique. *Psychol. Test Assess. Model.* **2015**, *57*, 577–594.
45. Van Buuren, S. *Flexible Imputation of Missing Data*, 2nd ed.; Chapman and Hall/CRC Press: New York, NY, USA, 2018.
46. Schafer, J.L. *Analysis of Incomplete Multivariate Data*; Chapman & Hall/CRC: New York, NY, USA, 1997.

47. Collins, L.M.; Schafer, J.L.; Kam, C.M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol. Methods* **2001**, *6*, 330–351. [CrossRef] [PubMed]
48. Graham, J.W.; Olchowski, A.E.; Gilreath, T.D. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci.* **2007**, *8*, 206–213. [CrossRef]
49. Enders, C.K. *Applied Missing Data Analysis*; The Guildford Press: New York, NY, USA, 2010.
50. Van Buuren, S.; Groothuis-Oudshoorn, K. Mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67. [CrossRef]
51. Lipsey, M.W.; Puzio, K.; Yun, C.; Hebert, M.A.; Steinka-Fry, K.; Cole, M.W.; Roberts, M.; Anthony, K.S.; Busick, M.D. *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms*; National Center for Special Education Research, Institute of Education Sciences (IES), U.S. Department of Education: Washington, DC, USA, 2012.
52. Kuhfeld, M.; Soland, J. The Learning Curve: Revisiting the Assumption of Linear Growth across the School Year. Available online: https://www.edworkingpapers.com/sites/default/files/ai20-214.pdf (accessed on 23 June 2020).
53. Bangert-Drowns, R.L.; Hurley, M.M.; Wilkinson, B. The Effects of school based writing-to-learn interventions on academic achievement: A meta-analysis. *Rev. Educ. Res.* **2004**, *74*, 29–58. [CrossRef]
54. Araya, R.; Jiménez, A.; Aguirre, C. Context-Based personalized predictors of the length of written responses to open-ended questions of elementary school students. In *Modern Approaches for Intelligent Information and Database Systems*; Sieminski, A., Kozierkiewicz, A., Nunez, M., Ha, Q., Eds.; Studies in Computational Intelligence; Springer International Publishing: Cham, Switzerland, 2018; Volume 769, pp. 135–146. [CrossRef]
55. Araya, R.; Plana, F.; Dartnell, P.; Soto-Andrade, J.; Luci, G.; Salinas, E.; Araya, M. Estimation of teacher practices based on text transcripts of teacher speech using a support vector machine algorithm. *Br. J. Educ. Technol.* **2012**, *43*, 837–846. [CrossRef]
56. Schlotterbeck, D.; Araya, R.; Caballero, D.; Jiménez, A.; Lehesvuori, S. Assessing Teacher's Discourse Effect on Students' Learning: A Keyword Centrality Approach. In *Addressing Global Challenges and Quality Education. EC-TEL 2020*; Alario-Hoyos, C., Rodríguez-Triana, M., Scheffel, M., Arnedillo-Sánchez, I., Dennerlein, S., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; Volume 12315, pp. 102–116. [CrossRef]
57. Uribe, P.; Jiménez, A.; Araya, R.; Lämsä, J.; Hämäläinen, R.; Viiri, J. Automatic Content Analysis of Computer-Supported Collaborative Inquiry-Based Learning Using Deep Networks and Attention Mechanisms. In *Proceedings of the Methodologies and Intelligent Systems for Technology Enhanced Learning, 10th International Conference, L'Aquila, Italy, 17–19 June 2020*; Vittorini, P., Di Mascio, T., Tarantino, L., Temperini, M., Gennari, R., De la Prieta, F., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 1241, pp. 95–105.